DOCUMENT RESUME

ED 449 209 TM 032 354

AUTHOR Kim, Sungsook C.
TITLE Investigating the Generalizability of Scores from Different Rating Systems in Performance Assessment.
PUB DATE 2000-04-00
NOTE 13p.; Paper presented at the Annual Meeting of the American Educational Research Association(New Orleans, LA, April 24-28, 2000).
PUB TYPE Reports - Research (143) -- Speeches/Meeting Papers (150)
EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS Concept Mapping; Foreign Countries; *Generalizability Theory; Middle School Students; *Middle School Teachers; Middle Schools; *Performance Based Assessment; *Scores
IDENTIFIERS South Korea

ABSTRACT

The generalizability of scores from different scales in performance assessment was studied. First, a concept map of teachers' and raters' perceptions about various scores and scales was constructed using multidimensional scaling analysis. Then, a generalizability study using a random, partially nested design was conducted to analyze the differences in the various rating systems. This study estimated the variance component of tasks, raters, and evaluative factor based on the scoring systems and determined the optimal number of grading conditions of each facet that maximized the generalizability coefficient. Data for the concept map were from questionnaires completed by about 218 middle school teachers in Korea. Data for the generalizability study were from two different scoring systems used to rate a report and presentation by each student in a middle school social studies class in Korea. The scores of 188 random samples used in the study were the interim scores of each factor before summing up a total score. Results show that the scoring of the performance task using the different rating systems was very consistent from rater to rater. However, the relatively large variance components suggested that the written report was rated differently across the different systems. Findings also suggest that when the student's report or presentation was being assessed, the generalizability of scores was enhanced by combining the ratings from more than one rater, mainly because this effectively increased the number of factors being evaluated. For ratings of performance, the generalizability coefficient increased considerably as the evaluative factors for the scoring standard became more specific. (Contains 19 references.) (SLD)

Reproductions supplied by EDRS are the best that can be made
from the original document.

TM

2000 AERA Annual meeting, New Orleans, Apr. 23-28

Session : Generalizability Theory Monday 4:05-5:35 ID No: D-54 Sponsors : D

Chair : Lukin, Leslie (Lincoln Public Schools)

Discussant : M. David Miller (University of Florida) & Behuniak, Peter

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

☐ This document has been reproduced as received from the person or organization originating it.

☐ Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

Investigating the generalizability of scores from different rating system in performance assessment

ED 449 209

Kim, Sungsook C.

Korea Institute of Curriculum and Evaluation(KICE)

S.C. Kim

I. Introduction

A direction of student evaluation in the classroom moves from paper-pencil test to performance-oriented assessment. When teachers assess students' achievement by various methods such as observation, portfolio, reports, etc., one problem that must be faced is ensuring the dependability of scores in different rating or grading system. A process of rating each performance especially contains a number of potential sources of error associated with raters, with tasks, with evaluative domains or factors, with different scales, and with their combinations.

Several studies that have applied generalizability theory (Cronbach, Gleser, Nanda, & Rajaratnam, 1972) to estimate the generalizability coefficient enables the researchers to analyze the influence of multiple sources of variance in performance assessment. In addition, effects of changing the number of observations in a single facet or two facet design on attaining a satisfactory level of the generalizability coefficient has been investigated extensively in the literature (Baxter, et al., 1992; Kim, 1998,1992; Lehmann, 1990; Croker, Llabre & Miller, 1988). A major contribution of G theory is that it allows the researcher to estimate the influence of sources of measurement error and increase the appropriate number of conditions of each facet so that the variations decrease. In other words, the researcher can compare the relative influence of each facet on a measure of the target assessment and estimate how many conditions of each facet are needed to attain a certain level of generalizability. However, effects of using different scores or scales in assessing students' performance have been investigated significantly in classroom. In other words, the results of students' achievement would turn out differently according to types of score or rating scale used. Lehmann (1990) stressed on the issue of scoring guide for written composition included sources of error related to intrarater and interrater effect.

Especially, scoring performance assessment contains possible errors such as rater

TM032354

BEST COPY AVAILABLE

disagreement, lack of objectivity, unclear rating guide, and changes over time in raters and environment. For example, reliability of essay item concerns the accuracy of measurement and the extent to which difference between the objects of the measurement can be dependably discriminated by the writing system. A process of scoring essay contains variations associated with raters (within a rater-group), with items, with evaluation domains or factors, with different rater groups, and with their combinations.

More specifically, for instance, we view a performance assessment as a sample of student achievement drawn from a complex universe defined by a combination of all possible task, occasion, raters and rating standards. We view the task facet to be representative of the content in a subject-matter domain. The occasion facet includes all possible occasions on which a rater would be equally willing to accept a score on the performance assessment. We view the rater facet as including all possible individuals who could be trained to score composition reliably. More importantly, the type of rating system includes different types of ratings or scores using in the classroom.

It is also true that teachers or raters perceive each rating or grading scale differently when they use those in assessing students' output. For example, most teachers in Korea use 100 point scores for marking the final grade whether they use a 3 point rating scale, a 5 point rating scale, or even a pass/fail system for evaluating student's performance during the semester. In other words, they multiplied each score by each weight of assessment assigned and added them up to 100 point score to make a final grade or rank. It can be a dangerous process because there is no evidence of using different rating scale for the similar output of assessment.

The purpose of the proposed study, therefore, which has been designed to improve upon the work of Frisbie & Waltman(1992), Abedi & Baker(1995) and Cronbach et al.(1997), is to compare the results of the generalizability of scores in different scales of performance assessment. First of all, a concept map of teachers/raters' perceptions about various scores and scales is constructed using multidimensional scaling analysis(Shephard et al., 1972; Tittle, et. al., 1996). Secondly, a generalizability study in a random, partially nested design is also conducted to analyze the variation in different rating systems. In particular, this G study provided two steps as follows: (1) estimating the variance component of tasks, raters, and evaluative factors based on different scoring system to compare the relative influence of each facet and (2) determining the optimal number of grading conditions of each facet that maximizes the generalizability coefficient.

The research questions related to the purpose were addressed as follows:

First, how do teachers or raters perceive relationships between different rating systems?

Second, is scoring student's performance generalizable across raters, different rating systems and tasks ?

a. What are the differences in the relative magnitudes of error variance due to raters, tasks, evaluative factors and interactions between these factors which influence on the generalizability of scores in performance assessment?

b. Does the generalizability coefficient improve by increasing the number of each facet? If so, how can we determine the optimal number of grading conditions of each facet that maximize the generalizability coefficient ?

II. Method

1. Data

The data used for the first research question in this study were based upon about 218 middle school teachers' questionnaire results conducted during May, 1999 in Seoul. The teacher questionnaire includes items related to similarities of scales enable researchers to focus on the multidimensional scaling of dissimilarity data as a way to construct objective scales of subjective attributes of items. Items in the questionnaire are pass/fail score, letter grading, 5 point rating score, 10 point rating score, 20 point score, 100 % corrected and percentile rank. The question related to how teachers think similar or different each rating scales or scores.

Another data used for the second research question in this study were based upon the results of essay and presentation for the performance assessment in social studies class conducted in June 1999 in one middle school in Seoul, Korea. Two different scoring systems were used to rate a report and a presentation of each student. A report was graded by two raters assigned to two different scoring systems, 5 point rating scales and 100 point scores, respectively. Each scoring system is composed to 5 evaluative factors supposed to be written or presented in the essay or presentation. One scoring system is assigned 25 points, 5 points per each evaluation factor, and another system is assigned 100 points, 20 points per each factor, therefore, the possible total score for essay and presentation is 250 points. A final score is summed from two raters, each score of domain is based on averages of two independent ratings. The scores of 188 random samples used in the study were interim scores of each factor before summing up a total score. Each presentation was rated as same as rating a report.

2. Analysis

1) MDS (Multidimensional Scaling)

MDS(Multidimensional Scaling) is designed to analyze distance-like data called dissimilarity data. MDS has its origins in psychometrics where it was proposed to help understand people's judgements or the similarity of members of set of a objects. Therefore, the purpose of applying MDS for the first research question is to construct a psychological map of the locations of scales or scores relative to each other from data that specify how different the scales or scores are.

Multidimensional scaling is accomplished by assigning observations to specific locations in a conceptual space(usually two- or three-dimensional) such that the distances between points in the space match the given dissimilarities as closely as possible. In many cases, the dimensions of this conceptual space can be interpreted and used to further understand the data. Multidimensional scaling can also be applied to subjective ratings of dissimilarity between objects or concepts. How do teachers perceive relationships between different rating or scores? If I have data from teachers indicating similarity ratings between different scales or scores, multidimensional scaling can be used to identify dimensions that describe raters' perceptions.

For each model of the MDS, optimally scaled data matrix, S-stress (Young's), stress (Kruskal's), RSQ, stimulus coordinates, average stress and RSQ for each stimulus (RMDS models) are calculated to make a conceptual map.

2) Generalizability theory

For the second research question, G theory uses the analysis of variance to provide estimates of scoring variation due to raters, tasks, evaluation factors and each source of error. By estimating the magnitude of the variance components, the sources of the greatest measurement error can be pinpointed. It is important to recognize that the purpose of a G study is to obtain estimates of variance components associated with the universe of admissible observations. More importantly, these estimates can be used to design efficient measurement procedures to provide information for making substantive decisions about objects of measurement, in various D studies. D study considers the specification of a universe of generalization, which is the universe to which decision maker wants to generalize in a D study. In particular, this study provided the generalizability of scores in performing as following two procedures : (1) estimating the variance components of raters, tasks, and evaluation factors to compare the relative influence of each facet and (2) determining the optimal numbers of grading conditions of each facet that maximize the generalizability coefficient.

The design addressing the questions included a three-facet generalizability study, ((p x (f : r) x t) design, with person(p) crossed with tasks and factor(f) within each raters(r). Since each rater use different rating system for evaluating same performance, evaluative factor is nested within each rater. In each ANOVA procedure, an estimate of the variance components corresponding to each factor and to each interaction between factors was calculated from the mean squares. The estimated variance components were then compared with one another for relative magnitudes in the results of different scoring systems, and the generalizability coefficient of interest, in which generalization of student's performance was over raters, tasks, evaluative factors and their interactions, was obtained.

According to the grading system, two raters assigned to each task scored all evaluative factors within each task, therefore, student(subject) effect crossed with raters, tasks, and factors. However, evaluation factors are nested within each raters. Therefore, the object of measurement is student(p: person) and sources of error include rater(r), task(t), and factor(f). The conditions of each facet can be defined as a sample of a complete set of conditions (i.e., fixed effect) or as the infinite set of conditions(i.e., random effect). For the design, students were considered to be a random effect because the students were chosen from possible classes. Raters were also treated as a random effect since raters were randomly chosen from eligible teachers in the middle school. In addition, performance tasks and evaluative factors also considered to be random effects. The data array of ((p x (f : r) x t) design can be displayed as shown as in [Figure 1].

	T1					T2				
	R1				r2	r1				r2
	f1	f2	f3	f4	f5	f6	f7	f8	f9	f10	f1	f2	f3	f4	f5	f6	f7	f8	f9	f10
S1	3	5	4	2	3	18	19	20	14	16	4	5	3	1	2	17	20	16	11	14
S2	2	4	3	2	2	10	16	07	08	10	3	4	4	2	1	07	10	13	12	10
S3	4	5	3	3	4	16	18	16	14	13	2	5	3	4	3	15	18	16	14	12
S4	.					.					.					.				

[Figure 1] An example data of G study p x (f : r) x t design

The design addressing the questions includes a three-facet generalizability study, p x (f : r) x t design, with evaluation factor(f) nested within rater(r) and crossed with tasks(t), and students (p) crossed with the other three factors. The variance of an observed score can be decomposed into nine variance components as follows:

$$\sigma_X^2 = \sigma_p^2 + \sigma_r^2 + \sigma_t^2 + \sigma_{f:r}^2 + \sigma_{pr}^2 + \sigma_{pt}^2 + \sigma_{rt}^2 + \sigma_{pf:r}^2 + \sigma_{tf:r}^2 + \sigma_{ptr}^2 + \sigma_{ptf:r,e}^2 \qquad (1)$$

In other words, the variance of the ratings can be partitioned into independent sources of variation due to difference between students, raters, tasks, factors within raters, their interactions, and the residual. In the notation for those components, the colon implied 'nested within', while two or three consecutive subscripts implied crossing of the effect. The focus of G study is on these variance components because their magnitude provides information about the sources of error influencing a measurement. In each ANOVA procedure, estimates of the variance components corresponding to each factor and to each interaction between factors were calculated from the mean squares. <Table 1> shows how to estimate each variance component from the analysis of variance.

<Table 1> G study 3 facet p x (f : r) x t design and the estimated variance components.

Rater(R):random($n_r \langle N_r \to \infty$), Factor(f):random($n_f \langle N_f \to \infty$), Task(t):random($n_t \langle N_t \to \infty$)

Effect	df	EMS
P	p-1	$\sigma_{ptf:r,e}^2 + n_t\sigma_{pf:r}^2 + n_f\sigma_{ptr}^2 + n_tn_r\sigma_{pf}^2 + n_rn_f\sigma_{pt}^2 + n_tn_f\sigma_{pr}^2 + n_tn_rn_f\sigma_p^2$
R	r-1	$\sigma_{ptf:r,e}^2 + n_p\sigma_{tf:r}^2 + n_t\sigma_{pf:r}^2 + n_f\sigma_{ptr}^2 + n_pn_t\sigma_{f:r}^2 + n_tn_f\sigma_{pr}^2 + n_pn_f\sigma_{tr}^2 + n_pn_tn_f\sigma_r^2$
T	t-1	$\sigma_{ptf:r,e}^2 + n_p\sigma_{tf:r}^2 + n_t\sigma_{ptr}^2 + n_rn_f\sigma_{pt}^2 + n_pn_f\sigma_{tr}^2 + n_pn_rn_f\sigma_t^2$
F:r	r(f-1)	$\sigma_{ptf:r,e}^2 + n_p\sigma_{tf:r}^2 + n_t\sigma_{pf:r}^2 + n_pn_t\sigma_{f:r}^2$
Pr	(p-1)(r-1)	$\sigma_{ptf:r,e}^2 + n_t\sigma_{pf:r}^2 + n_f\sigma_{ptr}^2 + n_tn_f\sigma_{pr}^2$
Pt	(p-1)(t-1)	$\sigma_{ptf:r,e}^2 + n_f\sigma_{ptr}^2 + n_rn_f\sigma_{pt}^2$
Tr	(t-1)(r-1)	$\sigma_{ptf:r,e}^2 + n_p\sigma_{tf:r}^2 + n_f\sigma_{ptr}^2 + n_pn_f\sigma_{tr}^2$
pf:r	r(p-1)(f-1)	$\sigma_{ptf:r,e}^2 + n_t\sigma_{pf:r}^2$

tf:r	r(t-1)(f-1)	$\sigma^2_{ptf,e} + n_p\sigma^2_{tf,r}$
Ptr	(p-1)(t-1)(r-1)	$\sigma^2_{ptf,e} + n_f\sigma^2_{ptr}$
ptr:r	r(p-1)(t-1)(f-1)	$\sigma^2_{ptf,e}$

p:student r:rater t:task f:factor

A generalizability coefficient is analogous to a classical reliability estimate except that distinct sources of measurement error are recognized and accounted for by the generalizable universe score. The student was the object of measurement in the scoring system, therefore, the variance component for students represented the universe score variance. The error variance included the variation related to interaction between raters and students, interaction between students and tasks, interaction between students and factors within raters, interaction among students, tasks and raters, and residual. A generalizability study can generate several coefficients, each corresponding to a different universe of conditions. The resulting estimated relative error variance and estimated generalizability coefficient can be expressed as follows:

$$\hat{\sigma}(\tau)^2 = \sigma(p)^2 \qquad (2)$$

$$\hat{\sigma}(\delta)^2 = \frac{\sigma^2_{pr}}{n_r} + \frac{\sigma^2_{pt}}{n_t} + \frac{\sigma^2_{ptf,r}}{n_f n_r} + \frac{\sigma^2_{ptr}}{n_t n_r} + \frac{\sigma^2_{ptf,r}}{n_t n_f n_r} \qquad (3)$$

$$\hat{\rho}(\delta)^2 = \frac{\sigma^2_{(\tau)}}{\sigma^2_{(\tau)} + \sigma^2_{(\delta)}} = \frac{\sigma^2_{(p)}}{\sigma^2_p + \left(\frac{\sigma^2_{pr}}{n_r} + \frac{\sigma^2_{pt}}{n_t} + \frac{\sigma^2_{ptf,r}}{n_f n_r} + \frac{\sigma^2_{ptr}}{n_t n_r} + \frac{\sigma^2_{ptf,r}}{n_t n_f n_r}\right)} \qquad (4)$$

The estimated error variance components were then compared for relative magnitudes. Each variance component contribute to several types of error variance for mean scores in a same D study p x (F : R) x T design and their contribution to such error variances can be reduced by increasing the D study sample size for each facet. The data were analyzed with GENOVA program developed by Brennan (1983) and computing for determining the number of grading conditions was completed manually.

III. Results and Interpretation

1. MDS results

The findings of the study present that the perception of each scores and scales was shown very meaningful. According to the final concept map, 3 point rating scales and 5 point rating scales were very similar to the coordinates from the metric analysis, on the other hand, 100 point score and percentage score were appeared some departures from other scales. The results supported that teachers would intend relative weights when he/she combined students' scores. The following <Table 2> and <Table 3> summarize the results of implementing MDS and the [Figure 2] plots the final concept map based on the data using the Euclidean distance

model.

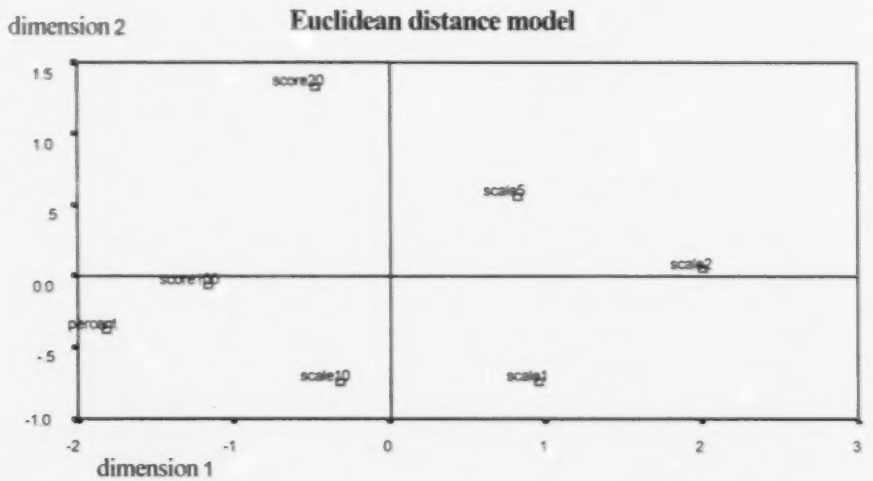
<Table 2> **Iteration history**

Iteration	S-stress	Improvement
1	.08876	
2	.06941	.01936
3	.06164	.00777
4	.05669	.00495
5	.05353	.00316
6	.05116	.00238
7	.05032	.00084

Final Stress = .04445 squared correlation(RSQ) = .98459

<Table 3> **Stimulus coordinates in 2 dimensions**

Stimulus number	variable	x (1)	y (2)
1	SCALE1	.9614	-.7433
2	SCALE2	2.0153	.0470
3	SCALE5	.8231	.5547
4	SCALE10	-.3260	-.7423
5	SCORE20	-.4818	1.3246
6	SCORE100	-1.1719	-.0668
7	PERCENT	-1.8201	-.3741



[Figure 2] **Plot of Stimulus coordinates for scales/scores**

The findings of the study present that the perception of each scores and scales was shown very meaningful. 3 point rating scales and 5 point rating scales were very similar to the coordinates from the metric analysis, on the other hand, 100 point score and percentage score were appeared some departures from other scales. The results supported that teacher would intend relative weights when he/she combined students' scores(Oosterhof, 1987).

2. G study results

The G study results showed that the variance component of universe score, e.g., the student's performance was relatively large (.047, 23.0%)(as shown in <Table 4>). Since the variation due to tasks(.042, 20.5%) and evaluative factors (.061, 29.8%) were large relative to the variation due to raters(.023, 11.2%), it is possible that each performance was scored differently on different tasks using different ratings. This indicates the generalizability of scorings is substantially influenced by types of scores or scales to evaluate each performance. On the other hand, the variance components of rater-related were zero or small, therefore, increasing the number of raters had approximately the small effect on changing the generalizability coefficient. The study presents the relative importance of raters, tasks, evaluative factors based on different scoring system in estimating the dependability of scores. Increasing the number of evaluative factors would be helpful to improve the generalizability coefficient more efficiently.

As a result of G study, generalizability of scores based on student's report indicates relatively high for the student and does not vary by raters. In other words, raters were well calibrated and varied little in their judgment of students' assessment. However, scorings report and presentation were affected by different evaluation factors. The D studies present the combinations of each facet to reach an acceptable generalizability coefficient and a decision maker can then examine the trade-off between the coefficient of generalizability and the total budget.

<Table 4> Results of G study 3 facet p × (f : r) × t design and proportion of the estimated variance components.

Domain(d):random($n_d \langle N_d \to \infty$), Factor(f):random($n_f \langle N_f \to \infty$), Rater(r):random($n_r \langle N_r \to \infty$)

Effect	df	σ_z^2	%
p	187	.047	23.0
r	1	.023	11.2
t	1	.042	20.5
f:r	8	.061	29.8
pr	187	.003	1.5
pt	187	.021	10.3
tr	1	.002	1.0
pf:r	1496	.0014	0.7
tf:r	8	.002	1.0
ptr	187	.0007	0.3
ptf:r, e	1496	.0014	0.7

Since the generalizability coefficient was calculated based on a combination of one task, one rater and one evaluative factor, it can be improved by trade-off between numbers of each facet needed to attain the specified acceptable level of generalizability, .80. The findings showed the effect of raters was less than that of evaluation factors, therefore, the optimal study for improving generalizability coefficient of the score would be based on having different combinations of fewer raters and more scoring standards.

As a result of D study, <Table 5> presented the trends of changing each estimated error variance and generalizability coefficient as increasing the numbers of task and factor when the numbers of raters are two. Increasing the number of evaluation domains or factors produced a better generalizability coefficient more efficiently than increasing the number of tasks. The G coefficient increased considerably as the number of evaluation factors increased. The level of generalizability of .80 was obtained at least with the combination of 2 tasks, 3 factors, and 2 tasks or 4 factors, 3 tasks and 3 factors. Therefore, the results indicated that the combination of different number of each facet was applied, having the combination of fewer tasks and more evaluative factors produced a better generalizability.

<Table 5> Results of D study 3 facets p×(R : F)×T design changing magnitude of error variance and G coefficient as increasing the number of task and evaluation factor (rater = 2)

SV		n_t	1	1	1	2	2	2	3	3
		n_f	1	2	3	3	4	5	3	4
P	σ_p^2		.047	.047	.047	.047	.047	.047	.047	.047
Pr	$\sigma_{pR}^2=\frac{\sigma_{pr}^2}{n_r}$		.0015	.0015	.0015	.0015	.0015	.0015	.0015	.0015
Pt	$\sigma_{pT}^2=\frac{\sigma_{pt}^2}{n_r}$		.021	.021	.021	.0105	.0105	.0105	.0035	.0035
pf:r	$\sigma_{pf:R}^2=\frac{\sigma_{pf:r}^2}{n_r n_f}$		.0007	.00035	.00023	.00023	.0002	.0001	.00023	.0002
Ptr	$\sigma_{pTR}^2=\frac{\sigma_{ptr}^2}{n_t n_r}$		.0007	.00035	.0001	.0001	.0000	.0000	.0000	.0000
Ptf:r,e	$\sigma_{pTf:R}^2=\frac{\sigma_{ptf:r}^2}{n_t n_r n_f}$		.0014	.0007	.00023	.00012	.000	.000	.000	.000
	$\hat{\sigma}(\tau)^2$		.047	.047	.047	.047	.047	.047	.047	.047
	$\hat{\sigma}(\delta)^2$		.0253	.0239	.02306	.01245	.0122	.0121	.0052	.0052
	$E\hat{\rho}^2$		.64	.66	.67	.79	.79	.80	.89	.90

n_f : number of evaluation factors, n_t : number of tasks

IV. Discussion

As I mentioned, when every teachers tries to evaluate students' activities and outcomes based on their performance, one problem that must be faced is assessing the dependability of scores in the grading system, despite performance assessment been an important role in emerging achievement. Once conceived as a sample of performance assessment from a complex universe, the statistical framework of generalizability theory can be brought to bear on the technical quality of achievement score. In terms of G theory, an assessment score or profile is but one of many possible samples from a large domain of assessments defined by the particular task, occasion, rater, rating standards. The theory focused on the magnitude of sampling variability due to items, rater, and so forth, and their combinations, providing estimates of the magnitudes of measurement error in the form of variance components.

Initially, technical evaluation of scoring student's outcomes focused primarily on the impact of rater sampling. With the complexity of students' performance assessment, the concern was that raters would be inconsistent in their evaluations. As our sampling framework suggests, defining the universe of generalization solely in terms of items and /or raters is limited. With complex of writing composition of essay, a student's achievement score may be impacted by several sources of sampling variability. Some are associated with generalizability, and others are associated with convergent validity. It, therefore, becomes important to estimate, simultaneously, as many potential sources of error- task, rater, occasion, and their interactions, etc.- and as many potential sources of grading system-standards and their interactions- as are feasible.

The results of the study showed that scoring of performance task using the different rating system was very consistent from rater to rater. However, the relatively large variance components of factor-related indicated that the written report was rated differently across different rating system. One possible explanation is that the 5 point rating system is scored more strictly than that of 100 point score. This findings suggested that, if a student's written report or presentation is being assessed, generalizability of scores was enhanced by combining ratings from more than one rater, but mainly because this effectively increased the number of factors evaluated. More studies (Shavelson & Webb, 1991; Engelhard, Jr., 1996) were confirmed the earlier findings that interrater reliability is not a problem, but task-sampling variability exists.

The study presented the relative importance of raters, tasks, factors, and their interactions in estimating the dependability of scores. Based on a result of relative size of each error variance, the study examined possible combinations of the conditions of each facet in order to determine the number of raters, evaluation factors and tasks that are needed to obtain an acceptable level generalizability for a measure. For this ratings of performance, the generalizability coefficient increased considerably as the evaluative factors for scoring standard are more specific. Also, because the raters within each task and rater-factor interaction variance components were zero or small, increasing the number of rater had approximately the same effect as increasing a corresponding number of factors in a rating system. However, if scoring written composition or presentation are used to provide an overal index of assessment of performance, there is a possible problem in validity associated with the use of specific standards for evaluation. Averaging the scores from several independent

ratings may result a final score of individual with high generalizability, but different score based on different tasks may vary considerably as the number of tasks increased.

While it is important to increase the level of generalizability, such is not always possible with limited resources. Maximizing reliability within a prespecified set of limited resources can be an important issue. A procedure was applied to determine the optimal number of grading process of each facet that can be used in a mixed design when the total budget is imposed. The D studies referenced (Lehmann, 1990; Ruiz-Primo, 1993) present the combinations of each facet to reach an acceptable generalizability coefficient, this decision can be obtained by using the procedure described above, and a decision maker can then examine the trade-off between the coefficient of generalizability and the total budget. Goldstein and Marcoulides (1991) have provided equations that can determine the optimal number of conditions that maximizes the generalizability coefficient for a fully-crossed random model. In a related study, Marcoulides and Goldstein (1992) have illustrated an example of a multivariate design in which one can choose the number of each facet for updating the generalizability coefficient when total budget is restricted. Therefore, the further study is to determine optimal measurement designs. As long as total budget for a research is known to the public, a simple procedure can be presented to determine the optimal number of observations and conditions of facet that maximize power and generalizability for fully crossed or partially nested, random model, or multifacet designs.

References

Abedi, J. & Baker, E.(1995). A Latent-variable modeling approach to assessing interrater reliability, topic generalizability, and validity of a content assessment scoring rubric. *Educational and Psychological Measurement*, 55(5), 701-715.

Baker, E. L and others (1996). Dimensionality and generalizability of Domain-independent performance assessments, *Journal of Educational Research*, 89(4), 197-205.

Baxter, G.P. , Shavelson, R.J.,Goldman, S.R., & Pine, J. (1992). Evaluation of procedure-based scoring for hand-on science assessment. *Journal of educational Measurement*, 29, 1-17.

Brennan, R. (1983). **Elements of Generalizability Theory**. Iowa City, 1A: The American College Testing Program.

Crick, J.E. & Brennan, R.L. (1983). **Mannual for GENOVA** : A generalized analysis of variation system. ACT bulletin, 43. The American College Testing Program. Iowa City, IA.

Croker, L., Llabre, M., & Miller, D.A. (1988). The generalizability of content validity ratings. *Journal of Educaitonal Measurement*, 25(4), 287-299.

Cronbach, L.J., Glesser, G.C., Nanda,H., & Rajaratnam,N. (1972). **The generalizability of behavioral measurement : The theory of generalizability for scores and profiles**. New York : John Wiley.

Cronbach, L.J., Linn, R.B., Brennan, R. & Haertel, E.(1997). Generalizability analysis for performance assessment of student achievement or school effectiveness. *Educational and Psychological Measurement*, 57(3), 373-399.

Engelhard,, Jr.(1996). Evaluating rater accuracy in performance. *Journal of Educational Measurement*, 33(1), 56-70.

Frisbie,D.A., & Waltman, K.K. (1992). Developing a personal grading plan. *Educational Measurement: Issues and Practices*, 11, 35-42.

Goldstein, Z. & Marcoulides, G.A. (1991). Maximizing the coefficient of generalizability in decision studies. *Educational and Psychological Measurement*, 51, 79-88.

Kim, Sungsook(1998). An Estimation of error-loss variation and dependability coefficient associated with cut-off scores. *Journal of Educational Evaluation (Korean)*, 11(1), 153-177..

Kim, Sungsook (1992). The dependability of student ratings of instructors across sections. ERIC: ED 346-158. Paper presented at the annual meeting of American Educational Research Association. San Francisco, CA.

Marcoulides, G.A. & Goldstein, Z. (1992). The Optimization of multivariate generalizability studies with budget constraints. *Educational and Psychological Measurement*, 52, 301-308.

Oosterhof, A.C. (1987). Obtaining intended weights when combining students' scores. *Educational Measurement: Issues and Practices*, 6, 29-37.

Ruiz-Primo, M.A., Baxter, G.P. & Shavelson, R. (1993). On the stability of performance assessments. *Journal of Educational Measurement*, 30, 41-53.

Shavelson, R., & Webb, N.M. (1991). **Generalizability theory: A Primer**. Newbury Park, CA: Sage

Shephard, A.K., Romney, & S.B. Nerlove(ed.). (1972). **Multidimensional Scaling: Theory and Applications in the Behavioral Sciences**. New York: Seminar Press

Tittle, C., Weinberg, S. & Hecht, D.(1996). Investigating the construct validity of scores from a measure of student perceptions about mathematics classroom activities using multidimensional scaling. *Educational and Psychological Measurement*, 56(4), 701-709.